# Performance Evaluation of Student Dataset using different Data Mining Algorithms

# Introduction

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. It also refers to *extracting or “mining” knowledge from large amounts of data*. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology and data dredging.

## Data Analysis and Knowledge Discovery

Data mining treated as synonym for another popular used term Knowledge Discovery from Data, or KDD. Data mining is a step in the knowledge discovery process; it is a process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories.

Data mining functionalities are used to specify the kind of patterns to be found in

data mining tasks.

In general, data mining tasks can be classified into two categories:

Descriptive and predictive. **Descriptive mining tasks characterize the general properties**

**of the data in the database**. In order to make predictions Predictive mining tasks perform inference on the current data

By data mining system we can generate thousands of patterns or rules but only small fraction of the patterns potentially generated would actually be of interest to any given user. A pattern is interested if it is (1) *easily understood* by humans, (2) *valid* on new or test data with some degree of *certainty*, (3) potentially *useful*, and (4) *novel*. A pattern is also interesting if it validates a hypothesis that the user *sought to confirm*. An interesting pattern represents knowledge[.[1]](#_References)

Here data mining is being used for analysis of Student Dataset as it would enable us to determine relationship among various attributes e.g. student age, citizenship, nationality, student university, Student Major and student level.

# Algorithms used for in-depth analysis of data

In current dataset Student data is having a large amount of Categorical Data; it is having 17 attributes; Date Of Birth, Age, Gender, Admission year, nation, nation-type, citizenship, residency, college, student\_major, Diploma description, cert\_average, cert date, student level, school and school city.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

In order to analyze and for extracting of useful information I would analyze data by various following data mining techniques and algorithms.

## Data Preprocessing

Today's databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources.

There are a number of data preprocessing techniques. *Data cleaning* can be applied to remove noise and correct inconsistencies in the data. *Data cleaning* (or *data cleansing*) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. So here I would be applying this technique on my student dataset and all missing values would be replaced by attribute average. For example in my dataset 30th record has no value under Nation, Nation-type and School City. After applying data cleaning method for replacing missing values school city would get average value e.g. sharjah as it is mean or average value. In section3 this function input and output is shown.

## Data Visualization

Its major part of data mining for analyzing and understanding data behavior. Here visualization and knowledge representation techniques are used to present the mined knowledge to the user.

## Mining Frequent Patterns

### What is frequent item-set mining?

Frequent pattern mining searches for recurring relationships in a given data set**.** Frequent item set mining leads to the discovery of associations and correlations among items in large transactional or relational data sets.

A typical example of frequent item set mining is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their “shopping baskets”. The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For instance, ***if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket***? Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space.

### Why to use Frequent Item-set Mining

Similarly,in given data set if I want to analyze Graduate student's behavior who belongs to college "Art, Humanities and Social sci" then what is the most student's citizenship, gender, Cert-Avg(High score) etc w.r.t their college.

I would like to get such information by using frequent item-set generation that what is likelihood value of graduate students with respect to their residency, citizenship, diploma description and gender as well.

### Efficient and Scalable Frequent Item set Mining Methods

Data can be analyze using frequent item-set mining by following algorithms

* 1. **Apriori**
  2. **FPGrowth**

**Apriori** is the basic algorithm for finding frequent item sets**.** Apriori generates candidate sets whereas FPGrowth uses specialized data structures (no candidate sets) so I would be using FPGrowth as it is more efficient and consume less memory than Apriori.

### FPGrowth

It found frequent item Set without candidate generation. It constructs a highly compact data structure (an *FP-tree*) to compress the original transaction database. Rather than employing the generate and-test strategy of Apriori-like methods, it focuses on frequent pattern (fragment) growth, which avoids costly candidate generation, resulting in greater efficiency.

### Generating Association Rules

Rule support and confidence are two measures of rule interestingness. Once the frequent item sets from transactions in a database *D* have been found, it is straightforward to generate strong association rules from them (where *strong* association rules satisfy both minimum support and minimum confidence). This can be done using Equation for confidence, which is here:

***Confidence (A🡪 B*) = *P* (*B* | *A*) = *support count* (*A* U *B*) /*support count* (*A*).[1]**

## **Classification**

Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large. I need analysis of student dataset that which school producing mostly graduated male/female students per year.

Data classification is a two-step process, as shown for the Student data in Figure 2.1. (The data are simplified for illustrative purposes. In reality, we may expect many more attributes to be considered.) In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or “learning from” a training set made up of database tuples and their associated class labels. A tuple, ***X***, is represented by an *n*-dimensional attribute vector, ***X*** = (*x*1, *x*2, : : : , *xn*), depicting *n* measurements made on the tuple from *n* database attributes, respectively, *A*1, *A*2, : : : , *An*.1 Each tuple, ***X***, is assumed to belong to a predefined class as determined by another database attribute called the class label attribute.[1]

|  |  |  |  |
| --- | --- | --- | --- |
| Training Data | | | |
| **CITIZENSHIP** | **Gender** | **SCHOOL** | **SCHOOL\_CITY** |
| UAE-Um Alquwain | M | UAE University | Al Ain |
| UAE-Sharjah | F | Ajman Univ of Science | AJMAN |
| UAE-Sharjah | M | Police Sciences Academy | SHARJAH |
| UAE-Sharjah | M | Police Sciences Academy | SHARJAH |
| UAE-Ras Alkhaimah | N | Islamic & Arabic Studies Coll | DUBAI |
| UAE-Abu Dhabi | M | Islamic & Arabic Studies Coll | DUBAI |
| UAE-Sharjah | N | Police Sciences Academy | SHARJAH |

Classification Algorithm

Classification Rules

If school is UAE University then school city is Al Ain.

If gender is M (male) and citizenship is UAE-Sharjah then School is police sciences academy.

If citizenship is UAE-Abu Dhabi and gender is M(male) then school is Islamic & Arabic Studies Coll

Fig 2.1(a)

Classification Rules

New Data

Test Data

|  |  |  |  |
| --- | --- | --- | --- |
| **Citizenship** | **Gender** | **School** | **School City** |
| UAE-Abu Dhabi | M | Ajman Univ of Science | Ajman |
| UAE-Abu Dhabi | F | UAE University | AL Ain |

(Gender M, UAE-Sharjah,Sharjah??

Police Sciences Academy

Fig 2.1(b)

**Fig 2.1** The data classification process: (a) *Learning*: Training data are analyzed by a classification algorithm. Here, the class label attribute is *school*, and the learned model or classifier is represented in the form of classification rules. (b) *Classification*: Test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples.

Here the data analysis task is classification, where a model or classifier is constructed to predict *categorical labels* for student's school such as Police Sciences Academy, Islamic & Arabic Studies Coll, Ajman Univ of Science, Police College In Abu Dhabi, University of Sharjah and UAE University and Kuwait University.

Suppose data analyst wants to predict what would be student major with respect to their gender then here student major would be class label and we can visualize student behavior well via decision tree as its results are shown in section 3.

## Evaluation of Classification Methods

Classification and prediction methods can be compared and evaluated according to the following criteria:

**Accuracy:** The accuracy of a classifier refers to the ability of a given classifier to correctly predict the class label of new or previously unseen data (i.e., tuples without class label information). The *confusion matrix* is a useful tool for analyzing how well your classifier can recognize tuples of different classes, The confusion matrix for six classes show in section 3.

### Classification by Decision Tree Induction Algorithms

Decision tree induction is the learning of decision trees from class-labeled training tuples.

There are many algorithms:

Hunt's Algorithm

CART

1D3, C4.5

**ID3** uses **information gain** as its attribute selection measure.

**C4.5** uses **Gain Ratio**

**CART** uses **Gini Index**

### Why Decision Tree?

Conducting analysis of decision making under uncertainty using decision trees serves several purposes.

* First, a decision tree is a visual representation of a decision situation (and hence aids communication).
* Second, the branches of a tree explicitly show all those factors within the analysis that are considered relevant to the decision (and implicitly those that are not).
* Third, and more subtly, a decision tree generally captures the idea that if different decisions were to be taken then the structural nature of a situation (and hence of the model) may have changed dramatically.
* Fourth, and arguably the most powerful, a decision tree allows for forward and backward calculation paths to happen and hence the choice of the correct decision to take (optimality of decision making, or optimal exercise if embedded real options) is made automatically.

## Cluster Analysis

Clustering partitions the large data-sets into groups according to their *similarity values*. A cluster is a collection of data objects that are *similar* to one another within the same cluster and are *dissimilar* to the objects in other clusters.

Cluster analysis can be used as a stand-alone tool **to gain insight into the distribution of data, to observe the characteristics of each cluster**, and to focus on a particular set of clusters for further analysis

Additional advantages of such a clustering-based process are that it is adaptable to changes and helps single out useful features that distinguish different groups.

1. Performance evaluation of data using K-Mean with different values of K(# of clusters) and would report precision, recall, accuracy and classification error
2. Performance evaluation of data using K-Mediods with different values of K(# of clusters) and would report precision, recall, accuracy and classification error
3. ***Evaluation of clusterings using the validation operators ClusterCentroidEvaluator and ClusterDensityEvaluator***
4. ***Evaluation of clusterings using the validation operators ClusterDensityEvaluator***

### Centroid-based clustering

### Density-based clustering